NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS
International College of Economics and Finance

Measuring the probability of informed trading in the cryptocurrency markets

Измерение вероятностей информированных торговых операций на рынках криптовалюты

Grigorii Kuzmin

July 17, 2021

Конкурс: Конкурс Банка России экономических исследований студентов и аспирантов вузов 2021 г

Автор: Кузьмин Григорий Иванович

Abstract: This paper focuses on economically and financially important problem of informed trading in the cryptocurrency markets. The metrics such as PIN and VPIN, more commonly estimated for stocks, are used for analysis and compared. The dynamics of the VPIN measure is qualitatively similar to that of the PIN, but it results much lower values, which could be partially explained by the market depth or too high order of Buy and Sell trades aggregation. Among other stylized facts the main finding is that more liquid cryptocurrencies have lower PIN.

Key words: PIN, VPIN, Probability of informed trading, Cryptocurrency markets, Order flow toxicity, Market microstructure

Contents

1 Introduction

A wide range of asset pricing literature focuses on identifying economic and other factors which determine the true value of financial assets. However, the majority of such models fail to account for specific factors, related to a range of characteristics of markets where assets are traded and ignore the mechanics of reaching the equilibrium. Financial microstructure is aimed at eliminating this gap and made great contribution to the understanding of financial market structure, its properties and functionalities.

In this research we focus on one specific class of financial microstructure models which spots the presence of information asymmetry by distinguishing between noise uniformed traders and informed traders, possessing private information. In particular we apply "PIN family" models (models, based on Glosten and Milgrom (1985) framework), enabling us to obtain the precise value of probability of insider trading. Efficient evaluation of information asymmetry contributes to markets regulation, making them more efficient. Moreover, such kind of metrics are worth including into asset pricing models, used by investors and financial institutions.

This paper is one of the first attempts to use "PIN family" models (PIN and VPIN), widely applied on stock markets, in order to evaluate the potential abuse of private information in cryptocurrency markets. Relatively new approach of using lower levels of trades aggregation for PIN estimation is applied. This makes estimation procedure more computationally intensive but provides higher precision. Lower aggregation order is found to be more efficient for evaluation of the models. Apart from other conclusions we observe that low volume cryptocurrencies have higher probability of insider trading which is consistent with Easley et al (1996) findings about stock markets. Furthermore, this study observes strong positive correlation between Buy and Sell trades, failed to be explained by traditional PIN framework, thus, an alternative Adj. Pin model (Duarte and Young, 2009) for future research is offered with several computational innovations.

The remainder is organised in the following way: Chapter 2 is a brief review of related literature is provided, Chapter 3 discusses main theoretical concepts used, Chap-

ters 4-5 provide overview of the package, created by the author and data used, and the final Chapters 6-10 present the analysis, further research potential and main conclusions.

2 Literature review

The core of financial microstructure works is based on the framework with three groups of agents: noisy and informed traders and the liquidity provider. These fundamentals, described by Kyle (1985) (introduced the order-driven model and the notion that prices are influenced by orders), and their further modifications, developed by such important papers as Glosten and Milgrom (1985), Hellwig (1980), Easley and O'Hara (1987) and others, are widely used in the recent research papers. There are many branches in microfoundations of financial markets, estimating the price impact of information, but the quantitive method of describing trading process, specifically, the method, finding the probability of informed trading, was firstly introduced by Easley, Kiefer, O'Hara and Paperman (1996). They considered an empirical fact that low volume stocks have larger bid-ask spreads, compared to those traded frequently. One of the possible reasons behind this phenomenon is private information. Their paper observes the difference in insider-trading between high and low volume stocks by estimating the direct measure of effect of informed trading (PIN). Easley and O'Hara conclude that active stocks have lower probability of informed trading.

Although PIN is a great tool for analysing the situation in the stocks markets and can be widely used for macroeconomic regression analysis, it has several limitations. PIN does not imply positive correlation between buyer and seller order flow which is observed empirically. Thus, the modified version of PIN called Adjusted PIN (*Adj* PIN) was introduced by Duarte and Young (2007). According to their paper, by allowing an event of symmetric Buy and Sell operations. *Adj* PIN matches the positive correlation between buy and sell order flow and provides more efficient estimate of probability of informed trading. They claim that in fact the original PIN apart from insider information estimation also accounts for illiquidity which is independent of private information. Still, both measures seem to identify the appearance of insider information.

There are also other further developments of PIN. Easley, López de Prado and

O'Hara (2012) introduced its modified version – VPIN (volume-synchronized probability of informed trading). This model simplifies the calculation as the intermediate estimation of PIN parameters is not required and, using the volume-clock paradigm, VPIN is assumed to provide good approximation for PIN. Moreover, this new metric serves as a proxy for the imbalance or "toxicity" of order flow. It predicted the famous "flash crash" which took place in May 6, 2010. However, VPIN is criticised for high dependence on the starting point of estimation in the dataset (Anderson and Bondarenko (2014)).

Another portion of critics is devoted to the fact that, according to several studies such as Aktas, de Bodt, Declerck, and Van Oppens (2007), Collin-Dufresne and Fos (2012) and others, PIN metric in different cases appears to be very low at the moment when the asymmetry of information is extremely high. This might be since the model does not fit the data or the imbalance in order flow is not enough to identify the private information events.

Finally, there exist several works that simplify the technical issues, related to the calculation of this metric. The PIN estimation requires the maximization of the likelihood function which includes factorials of numbers of Buy and Sell trades. The significant increase in these values decreases the feasible set of solutions or factorials become too large and cannot be computed. To avoid this bias, alternative methods of factorization were introduced by Easley et al (2010), Lin and Ke (2011), where they drop the large factorial terms from maximization problem as they become constant. Another problem associated is the choice of initial parameters' values in the optimisation problem. Yan and Zhang (2012) developed a special algorithm which helps to avoid bias in the final outcome. Finally, Gan (2015) introduced a clustering approach where the data is clustered into three groups: good news, bad news and no news via mean difference. It is shown that the resulting estimates are the best for the initial values of parameters in the likelihood maximization.[1]

[1]This literature review is based on G. Kuzmin (2020)

3 PIN (EHO)

In this section we are going to briefly discuss the PIN model, however, we do not consider the original version (EKOP) [Easley, Kiefer, O'Hara and Paperman (1996)] rather then its modification (EHO) [Easley, Hvidkjaer and O'Hara (2002)] as it is almost the same but with one small difference that is more applicable for our research and on which we will comment later in this section.

One of the stylized facts is that low volume stocks have larger bid-ask spreads, compared to those traded frequently. There are several reasons, explaining this fact: inventory or liquidity effect, market power and private information. This paper (EHO) observes the difference in insider-trading between high and low volume stocks by estimating the direct measure of effect of informed trading (PIN). They conclude that active stocks have lower probability of informed trading.

Model outline

There are two types of traders: informed traders who obtain private information and use it for speculative trading and noise traders, trading for liquidity or other exogenous purposes. Moreover, there is a market maker, setting bids and ask quotes according to buy and sell orders flow and estimating the probability of receiving orders from informed traders. At the beginning of each day there is an independent private information event which occurs with probability α. This event can be bad (negative signal) with probability δ and good (positive signal) with probability *(1-δ)*. Defining the arrival rate of uninformed buyers and sellers and informed traders is distributed by Poisson process as ϵ_b, ϵ_s and μ respectively, on the day with positive signal the total *buy order flow* is $\mu + \epsilon_b$ and the total *sell order flow* is ϵ_s. On the day with negative signal everything is vice versa. (See the Figure 1 below)

Figure 1: Trading process tree.

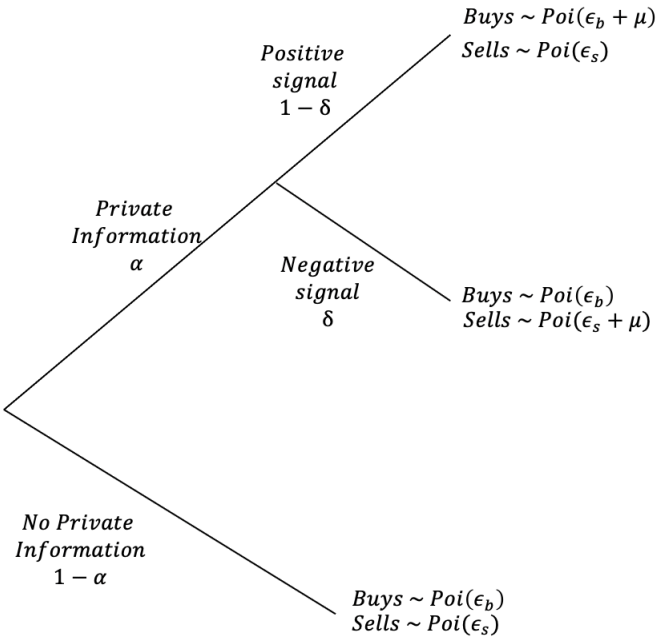


Figure 1: **Trading process tree.** This diagram represents the trading mechanics, described on the previous page, where α, δ, ϵ_b, $epsilon_s$ and μ stay for probability of private information event, probability of negative signal, rate of uninformed buy and sell operations rates and informed trade arrival respectively.

Using homogenous Poisson processes, the following Likelihood function is derived:

$$\begin{aligned} L(\Theta|B,S) = (1-\alpha) \times e^{-\epsilon_b} \times \frac{\epsilon_b^B}{B!} \times e^{-\epsilon_s} \times \frac{\epsilon_s^S}{S!} \\ + \alpha \times \delta \times e^{-\epsilon_b} \times \frac{\epsilon_b^B}{B!} \times e^{-(\mu+\epsilon_s)} \times \frac{(\mu+\epsilon_s)^S}{S!} \\ + \alpha \times (1-\delta) \times e^{-(\mu+\epsilon_b)} \times \frac{(\mu+\epsilon_b)^B}{B!} \times e^{-\epsilon_s} \times \frac{\epsilon_s^S}{S!} \end{aligned} \quad (1)$$

where $\Theta = (\alpha, \delta, \mu, \epsilon_b, \epsilon_s)$ is a vector of parameters and $(1-\alpha),\ \alpha \times \delta,\ \alpha \times (1-\delta)$ are no news, bad news, good news trading days respectively, while B and S are total Buy and Sell operations per day.

Formulating the maximization problem for t trading days is similar to the product of daily Likelihoods functions. Taking *log* (monotonic transformation) makes this

equivalent to the sum of daily Log Likelihood functions:

$$V = \prod L(\Theta|M) = \sum logL(\Theta|M)$$

where $M = ((B_1,S_1),...,(B_n,S_n))$ is vector of Buy and Sell orders.

Finally, having solved the maximization problem, we obtain the optimal parameters values which are used for PIN calculation. PIN (*probability of informed trading*) is calculated as ratio of expected informed arrival rate to expected total arrival rate:

$$PIN = \frac{\alpha \times \mu}{\alpha \times \mu + \epsilon_b + \epsilon_s} \tag{2}$$

Formula 2 takes into account both insider and noise trading and beliefs of liquidity provider. For instance, if there are only informed trades, based private information, ($\epsilon_s = \epsilon_b = 0$) then *PIN=1* and there is a wide bid-ask spread. Considering the case without insider trading ($\alpha = 0$), the *PIN=0* is obtained and there is no spread. Here we can observe the main distinction in the approach in the EHO model presented above and the original EKOP. In the EKOP model there is no differentiation between uninformed buyers and seller, they are assumed to act at the same rate $\epsilon_s = \epsilon_b = \epsilon$. However, in EHO, which we use in this paper, liquidity buyers and seller participate with uniques rates ϵ_b and ϵ_s respectively.

3.1 Factorization techniques and initial parameters

The original factorization in Likelihood function, demonstrated in the previous section, proved to be inefficient for highly liquid securities as it was not possible to estimate the factorial of high number. Even the log form does not solve this problem although it is supposed to decrease the value of a factorial. Thus, several alternative methods were introduced by Easley, Hvidkjaer and O'Hara. (2010) and Lin and Ke (2011).

3.2 Easley, Hvidkjaer and O'Hara. (2010) factorization

Making several rearrangements and eliminating the constant term does not affect the maximization the following Log Likelihood function for T trading days is derived:

$$Log\ L(\Theta|(B_t,S_t)_{t=1}^{T}) = \sum_{t=1}^{T}[-\epsilon_b - \epsilon_s + M_t(\ln x_b + \ln x_s) + B_t \ln(\mu+\epsilon_b) + S_t \ln(\mu+\epsilon_s)] + \\ + \sum_{t=1}^{T} \ln[\alpha(1\delta)e^{-\mu} x_s^{S_t-M_t} x_s^{B_t-M_t} x_b^{-M_t} + \alpha\delta e^{-\mu} x^{B_t-M_t} x_s^{-M_t} + \\ + (1-\alpha) x_s^{S_t-M_t} x_b^{B_t-M_t}] \quad (3)$$

where $M_t = min(B_t,S_t) + \frac{max(B_t,S_t)}{2}$, $x_s = \frac{\epsilon_s}{(\mu+\epsilon_s)}$, $x_b = \frac{\epsilon_b}{(\mu+\epsilon_b)}$. This modification solves the computational problems related to highly liquid stocks with large number of Buy and Sell operations.

3.3 Lin and Ke (2011) factorization

Another solution to the same problem was introduced by Lin and Ke (2011). They introduced the following modification for a Log Likelihood for one trading day:

$$L(\Theta|B_t,S_t) = \ln[\alpha\delta exp(e_{1i} - e_{maxi}) + \alpha(1-\delta)exp(e_{2i} - e_{maxi}) + (1-\alpha)exp(e_{3i} - e_{maxi})] + \\ + B_t \ln(\epsilon_b + \mu) + S_t \ln(\epsilon_s + \mu) - (\epsilon_b + \epsilon_s) + e_{maxi} - \ln(S_t! B_t!) \quad (4)$$

where $e_{1i} = -\mu - B_t \ln(1 + \frac{\mu}{\epsilon_b})$, $e_{2i} = -\mu - S_t \ln(1 + \frac{\mu}{\epsilon_s})$, $e_{3i} = -B_t \ln(1 + \frac{\mu}{\epsilon_b}) - S_t \ln(1 + \frac{\mu}{\epsilon_s})$ and $e_{maxi} = max(e_{1i},e_{2i},e_{3i})$. The constant term in the end can be dropped as does not affect the maximization. Practically, this method proves to be more efficient than the one by Easley et al (2010) as can be applied to even larger numbers, which its alternative fails to estimate. Moreover, it is claimed to reduced the downward bias present in the traditional method.

3.4 Initial parameters: Yan and Zhang (2012) algorithm

Yang and Zhang (2012) finds that Likelihood maximization in PIN model often results into boundary solutions, which in turn creates bias. In order to avoid it, special algorithm was developed for choosing the initial parameters values, used in the optimization process. Firstly, they derive the marginal expected values of B (total buy trades)and S (total sell trades):

$$E(B) = \alpha(1-\delta)\mu + \epsilon_b$$

$$E(S) = \alpha\delta\mu + \epsilon_s$$

Using these two equations above, authors are able to set the initial values for parameters:

$$\alpha_0 = \alpha_i,\ \delta_0 = \delta_j,\ \epsilon_b^0 = \gamma_k \bar{B} \tag{5}$$

$$\mu_0 = \frac{\bar{B} - \epsilon_b^0}{\alpha_0(1-\delta_0)} \tag{6}$$

$$\epsilon_s^0 = \bar{S} - \alpha^0\delta^0\mu^0 \tag{7}$$

where α_i,δ_j,γ_k take one of equally-distanced values (0.1,0.3,0.5,0.7,0.9) at a time (there are 125 possible combinations), $\bar{B}$ and $\bar{S}$ are estimators of expected values of B and S respectively. Thus, we arrive at the following procedure: firstly, we run maximization for the 125 sets of the possible values, excluding those where ϵ_s is negative ($\epsilon_s < 0$). Secondly, if all solutions are on the boundary, we choose the one with the highest value of the Likelihood function, otherwise we exclude them and choose the one among non-boundary, using the same approach.

3.5 VPIN

Introduced by Easley et al (2012), VPIN is aimed at estimating the order flow toxicity, where passive orders are filled more quickly than they should and vice versa filled slowly when they should be filled quickly. VPIN metric is alternative measure to PIN, calculated via volume based approach. In contrast to original PIN (EHO) model, for VPIN estimation there is no need in evaluating the unobservable parameters which facilitates the estimation procedure. Moreover, VPIN uses trade time instead of clock time, which is supposed to increase the accuracy.

VPIN estimation is divided into three steps. Firstly, all trade volumes are aggregating volume (time) bars. The common size, proposed by Easley at al. (2012), is 1 minute time interval, so all the trades within 1-min time bar summed up. Furthermore, the difference between the opening and closing price within every interval is estimated. (See Table 1)

Table 1: Time Bar aggregation

Time bar	Volume	Price change
12:00:01-12:01:00	10000	1$
12:01:01-12:02:00	13450	-0.5$
...	...	...

Secondly, the volume bucketing procedure is applied. The whole day trading volume is divided by 50 and this result is the volume bucket size (VBS) each separate bucket V_i. Thus, there are 50 buckets in total. Using the volume values from time bars calculated in the first step, we fill the bucket until we get VBS and if there are some trades left in the particular time bar, they are used to fill the next bucket.

The third step is the identification of the Buy and Sell trades. The order flow data is usually homogenous and Buy or Sell trades are not identified beforehand. That is why, special methods were developed such as *Lee-Ready algorithm* (discrete classification widely applied for PIN (EHO) model). In VPIN model authors introduce the

continuous approach and define trade direction in probabilistic terms:

$$V_i^{Buy} = V^i \times \Phi\left(\frac{P_i - P_{i-1}}{\sigma_{\triangle P}}\right)$$

$$V_i^{Sell} = V^i \times (1 - \Phi\left(\frac{P_i - P_{i-1}}{\sigma_{\triangle P}}\right))$$

where V_i is VBS of bucket i, $\sigma_{\triangle P}$ is the standard deviation of price differences between all time bars and Φ is c.d.f of rather normal or student's t-distribution. Such technique helps to increase the proportion of Buy trades when the price increases.

Finally, we use the classified volumes for estimating VPIN metric:

$$VPIN = \frac{\sum |V_i^{Buy} - V_i^{Sell}|}{nV_i} \approx \frac{\alpha \times \mu}{\alpha \times \mu + \epsilon_b + \epsilon_s}$$

4 R package "pin family"

Empirical results are obtained via author's own R package, called "pin family" without any usage of already existing packages for estimation of PIN-family models.[2]. This new package enables us to estimate pin and vpin models, allowing to change the underlying algorithms or specifications.

4.1 PIN code

```
> pin(data = ... , factor = ... , aggr = ...)
```

This function estimates the model's parameters $\alpha, \delta, \mu, \epsilon_b, \epsilon_s$ and the resulting PIN value. It has three variables: *data*, *factor* and *aggr* which stand for Buy and Sell trades matrix, type of factorisation and data aggregation level. As for factorisation in the likelihood function, two specifications are offered: the one, provided by Lin and Ke (2011) (*factor* = *"LK"*) or the one, introduced by Easley et al (2010) (*factor* = *"EHO"*), respectively. Data can be aggregated on five levels: *by day* (aggr = 1d), by 2 hours (aggr = 2h), *by 1 hour* (aggr = 1h) and *by 30 minutes* (aggr = 30m).

For example, for XBTUSD(*Bitcoin*) we have the following hourly aggregated trades Buy and Sell trades matrix for 01.01.2021:

Buy	Sell
172,303,511	126,078,952
50,114,584	73221068
46,019,359	57,940,714
52,182,277	59,560,492
67,292,451	58,125,139
77,573,158	75,757,555
137,938,418	125,920,994
72,403,922	88,312,776
60,720,464	66,775,533
99,699,238	110,206,645
54,798,616	48,836,721
33,667,222	23,411,475

[2]If you are interested in getting access to author's package, contact gikuzmin@edu.hse.ru

```
  > pin(data, "LK", "2h")

  [,1]        [,2]
[1,] "alpha"     "0.250090559840515"
[2,] "delta"     "0"
[3,] "mu"        "79450160.9340288"
[4,] "epsilon_b" "57196894.6184804"
[5,] "epsilon_s" "76179008.649304"
[6,] "PIN_value" "0.12965938490989"
```

4.2 VPIN code

```
> vpin(data = ... , buckets = ... , aggr = ...)
```

Function for VPIN model has three arguments: *data, buckets and aggr* that stand for the HFT monthly data, number of buckets used and level of aggregation, respectively. As for buckets number, the default value of 50 is taken, while default value of aggregation level is 1 minute. If we consider the same XBTUSD ticker as above for 01.01.2021:

```
   > vpin(data, 50, "1min")


      date           VPIN       Initial bucket Final bucket
1  2021-01-01     5.988502e-05          1           50
2  2021-01-01     3.355495e-03         51          100
---
30 2021-01-01     5.393869e-04        1451         1500
31 2021-01-01     7.571016e-04        1501         1550
```

5 Data

In this paper we analyze over 12 tickers for a period between 2018 and 2021 years. Relevant high frequency data is used from *BitMex* cryptocurrency market. Although it might not be the largest existing crypto exchange, but is commonly known to be the leading "indirect" volume mover as it trades mainly through leverage or margin contracts. Moreover, *BitMex* accounts for all types of orders, including hidden ones.
We divide cryptocurrencies into two groups in terms of their liquidity: ***Liquid*** and ***Not Liquid***.
In order to distinguish between these two, *Amihud Illiquidity measure* is used (Amihud, Yakov, 2002).

$$\text{Amihud Illiquidity} = \frac{1}{T}\sum_{t=1}^{T}\frac{|r_t|}{V_t}$$

where r_t, V_t are return and volume at period t, respectively. The measure itself has no economic interpretation, still it enables us to distinguish between assets in terms of liquidity. The lower is the Amihud Illiquidity measure, the more liquid is an asset. As a result a result we can form two groups of tickers in terms of Liquidity:

Liquid	Not Liquid
Bitcoin (XBT)	Litecoin (LTC)
TRON (TRX)	ESO.IO (EOS)
Ethereum (ETH)	Bitcoin Cash (BCH)
Cardano (ADA)	Chainlink (LINK)
Ripple (XRP)	Polkadot (DOT)

This distribution changes slightly throughout the considered years as trading data for some of tickers might be missing for a particular year or for another period of time. For Amihud Illiquidity estimations see **Appendix**.

The analysis of such a large amount of HFT data (>100 GB) is very computationally intensive and required over 15 computers estimating simultaneously to speed up the process.

6 PIN empirics

One of the first things of interest is the effect on PIN value in case of using different aggregation techniques. The traditional PIN estimation approach implies the usage of total Buy and Sell number of trades. However, in terms of high-frequency data it might seem to be not efficient for a number of reasons. Firstly, such an aggregation contains very limited amount of information as we account only for the aggregate amount of data for the whole day. Secondly, these values may be too noisy, as due to high liquidity there is a large number of uninformed, noise traders and using total daily trades might be not enough to correctly spot potential presence of insider activity. Thus, the application of more frequent aggregation such as hourly aggregation or 2-hours aggregation might increase the precision of PIN estimation, however, at a cost of much greater computational intensity. In this paper we directly compare four types of aggregation: *by day, by 2 hours*, *by 1 hour* and *by 30 minutes*.

In **by day** aggregation (d. aggr) we simply use a vector of total Buy and Sell operations for that particular day as an input for further ML estimation:

$$data = c(Buy, Sell)$$

$$Buy = \sum buy\ trades\ at\ day\ i$$

$$Sell = \sum sell\ trades\ at\ day\ i$$

However, we introduce relatively new approach which is aggregating by hour (1h aggr.). We obtain a list of two vectors with 24 hourly aggregated Buy and Sell trades, respectively. We obtain daily PIN value as a result of maximisation of the sum of 24 likelihood functions. We can interpret such technique as increasing the number of potential insider information signals from one per day to every single hour.

$$data = list(c(Buy_1, Sell_1), ..., c(Buy_{24}, Sell_{24}))$$

$$Buy_i = \sum buy\ trades\ at\ hour\ i$$

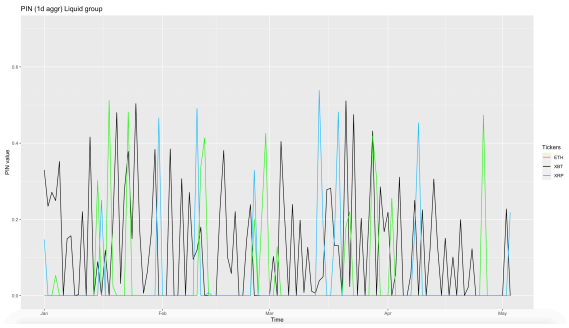
$$Sell_i = \sum sell\ trades\ at\ hour\ i$$

The remaining two approaches of aggregation by 2 hours (2h aggr.) and by 30 minutes (30m aggr.) are performed in the same way, but at different frequencies.

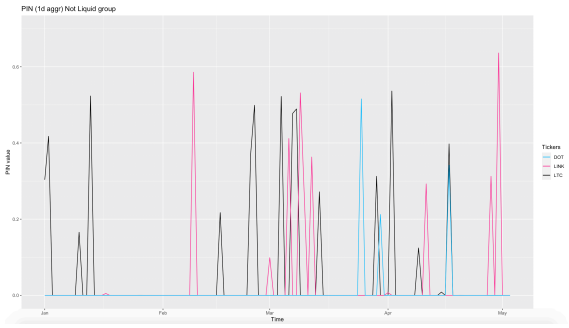
6.1 PIN series I (aggregation by day)

Intuitively, even before precise estimation we expect these series to be volatile with infrequent, sharp peaks. Moreover, they should stay in the area of zero most of the time. This should be due to the fact that cryptocurrency markets are extremely liquid with millions or even billions of operations per day and aggregated values of total Buy and Sell trades within one trading day contains a lot of noise, making the final PIN value not very informative.

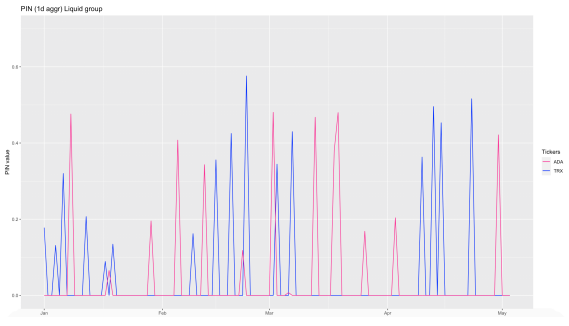
Figure 2: PIN (1d aggr) Liquid group vs Not liquid group (2021)



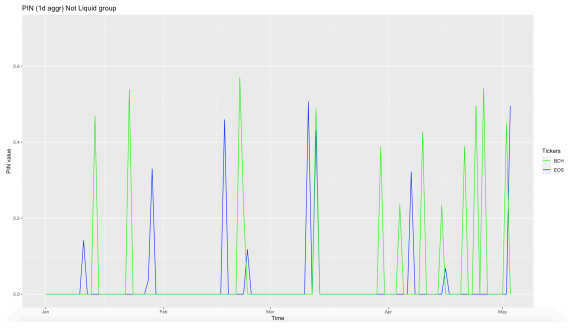
(a) Liquid group (XBT, XRP, ETH)



(b) Not Liquid group (DOT, LINK, LTC)



(c) Liquid group (ADA, TRX)

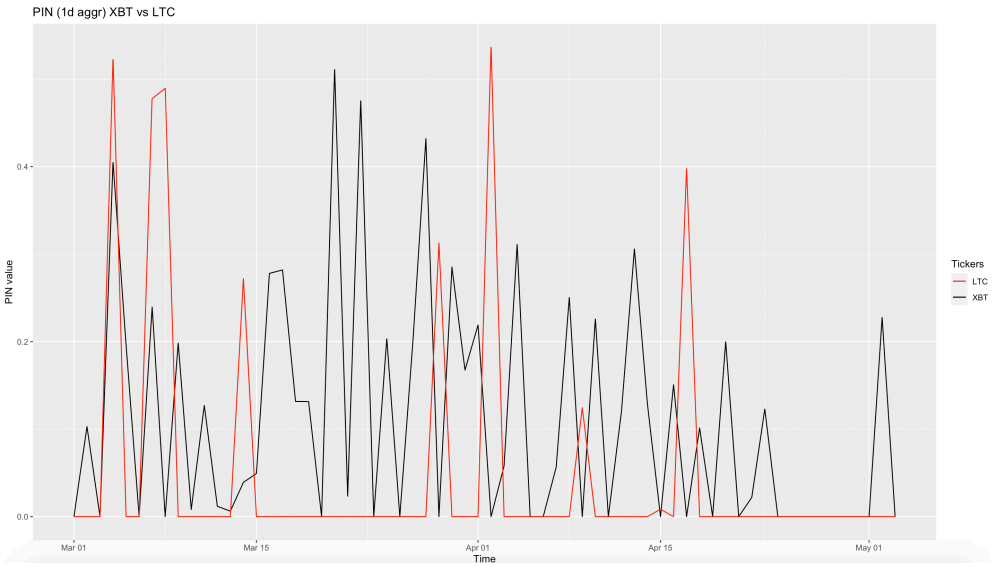


(d) Not Liquid group (BCH, EOS)

The graphical representation 4 of 2021 PIN time series above reflects our expecta-

tions and series indeed are quite volatile and take many zero values, creating "saw" shape. At the first glance, it is clear that more liquid cryptocurrencies appear to have higher proportion of days with PIN larger than zero at this level of aggregation. The potential reason behind this is directly related to liquidity. Liquid cryptocurrencies are traded more and some public or private information events, related to them, occur more frequently, compared to less liquid cryptocurrencies. That is why, we observe smoother fluctuation in PIN values of liquid group. However, in case of information event Not Liquid group appears to have relatively higher PIN than Liquid one, on average. This is consistent with the Easley et al (1996) that observed that illiquid stocks appear to have higher PIN than liquid ones.

These features become even more evident at lower scale, if we directly compare two different pairs of one liquid cryptocurrency to illiquid one.



(a) XBT (liquid) vs LTC (illiquid)

Average parameter values (1d aggr) for Liquid group (2018-2021)

Ticker	alpha	delta	mu	epsilon_b	epsilon_s
XBT	0.53	0.60	41,668,200,616	1,145,339,086	1,095,131,458
TRX	0.56	0.44	417,028,887	86,237,806	91,674,393
ETH	0.61	0.56	287,381,267	72,144,088	71,984,088
ADA	0.58	0.56	217,284,963	33,222,285	35,663,244
XRP	0.58	0.58	188,475,094	20,737,092	21,433,905

Average parameter values (1d aggr) for Not Liquid group (2018-2021)

Ticker	alpha	delta	mu	epsilon_b	epsilon_s
LTC	0.50	0.61	8,474,374	495,570	503,801
EOS	0.44	0.52	12,026,733	578,349	595,234
BCH	0.47	0.62	6,665,761	354,562	357,700
LINK	0.19	0.24	1,766,414	72,770	74,980
DOT	0.12	0.16	88,727	5,282	5,191

From two tables above we can infer that liquid group has higher probability of information event (α) which is consistent with our intuition for the smoother shape of PIN series of Liquid cryptocurrencies. The probability of positive signal seems to be also higher for frequently traded cryptocurrencies, still these two patterns are offset by extremely much larger market depth in case of liquid group. Liquid crypto's ϵ_b and ϵ_s are much higher then the ones of illiquid cryptocurrencies, indicating that for liquid group the market is very deep and the scale if noninformation-linked/liquidity trading is large. This is also the reason why in case of presence of insider information not liquid cryptos show higher PIN values due to the fact that less actively traded tickers are riskier as it is more probable that a trade is conducted by an insider.

Thus, we can summarise the following PIN (1d aggr.) **stylized facts**:

1) Liquid cryptocurrencies (at PIN 1 day aggregation) seem to fluctuate more smoothly and have larger proportion of non-zero values

2) Liquid cryptocurrencies (at PIN 1 day aggregation) tend to have higher probability of information event, but markets for them are much deeper.

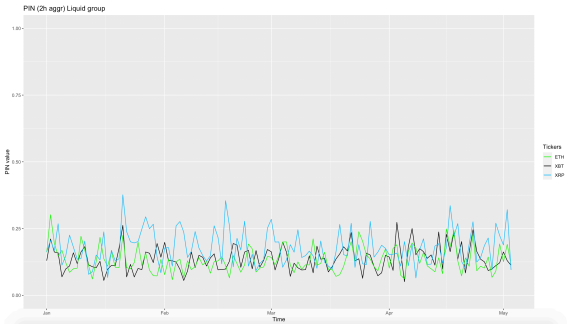
3) Illiquid cryptocurrencies (at PIN 1 day aggregation) have more zero-values apparently due to the absence of any information events, related to them, however, in case of presence of informational event they appear to have **higher** PIN values, compared to liquid cryptocurrencies.

6.2 PIN series II (aggregation by 2 hours)

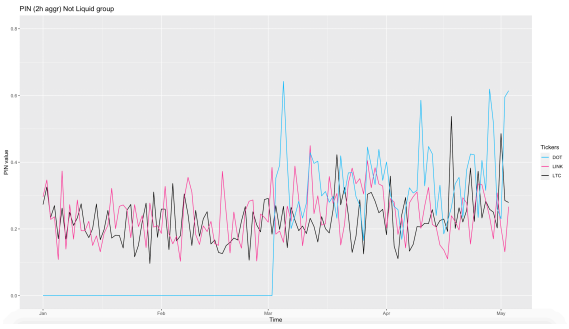
From lower level of aggregation we expect much smoother PIN series as such Buy and Sell trades vectors contain much more information, compared to just two sums. They might reflect different effects within one day that may enable us to capture insider trading more efficiently.

In order to contrast previous type of aggregation, let us consider the same 2021 year period. As expected, these PIN series are more monotonous and fluctuate much more evenly. However, the difference in PIN values between two groups decreased: although DOT, LINK and LTC appear to have higher PIN values than XBT, XRP and ETH, on average, PIN-s for BCH and EOS are close to the ones of ADA and TRX.

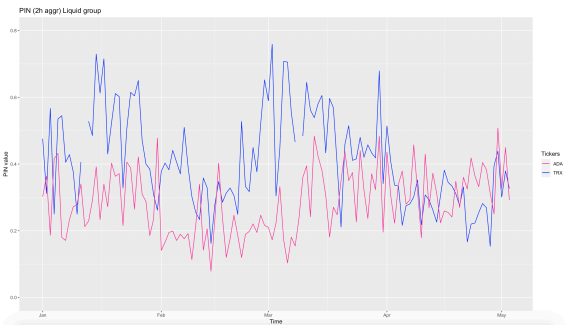
Figure 4: PIN (2h aggr) Liquid group vs Not liquid group (2021)



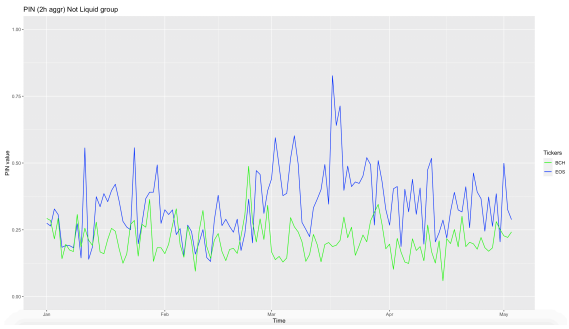
(a) Liquid group (XBT, XRP, ETH)



(b) Not Liquid group (DOT, LINK, LTC)



(c) Liquid group (ADA, TRX)



(d) Not Liquid group (BCH, EOS)

From the tables and the graphs it is clear the all the **stylized facts** for PIN (1d aggr) are valid for PIN (2h aggr). However, the PIN (2h aggr) series contain more information about the underlying trades and, thus, are much smoother and appear to indicate that there is permanent non-zero probability of informed trading.

Average parameter values (2h aggr) for Liquid group (2018-2021)

Ticker	alpha	delta	mu	epsilon_b	epsilon_s
XBT	0.30	0.55	132,973,354	84,041,847	79,822,229
TRX	0.30	0.45	21,464,848	5,042,834	5,181,746
ETH	0.31	0.56	8,712,111	5,138,260	5,072,344
ADA	0.34	0.55	7,127,068	2,071,106	2,061,817
XRP	0.32	0.57	3,927,811	1,370,968	1,386,439

Average parameter values (2h aggr) for Not Liquid group (2018-2021)

Ticker	alpha	delta	mu	epsilon_b	epsilon_s
LTC	0.35	0.56	77,555	33,413	29,936
EOS	0.29	0.45	142,368	34,598	34,109
BCH	0.35	0.56	51,264	24,205	23,098
LINK	0.14	0.23	10,801	4,966	4,641
DOT	0.09	0.13	1,670	244	234

6.3 PIN series III-IV (1h and 30 min aggregation)

These lower types of aggregation follow the same pattern as previous two (PIN 1d and PIN 2h), so we will focus on rather comparing them to each other.

Figure 5: PIN (1h) vs PIN (30 min) for XBT (2021)

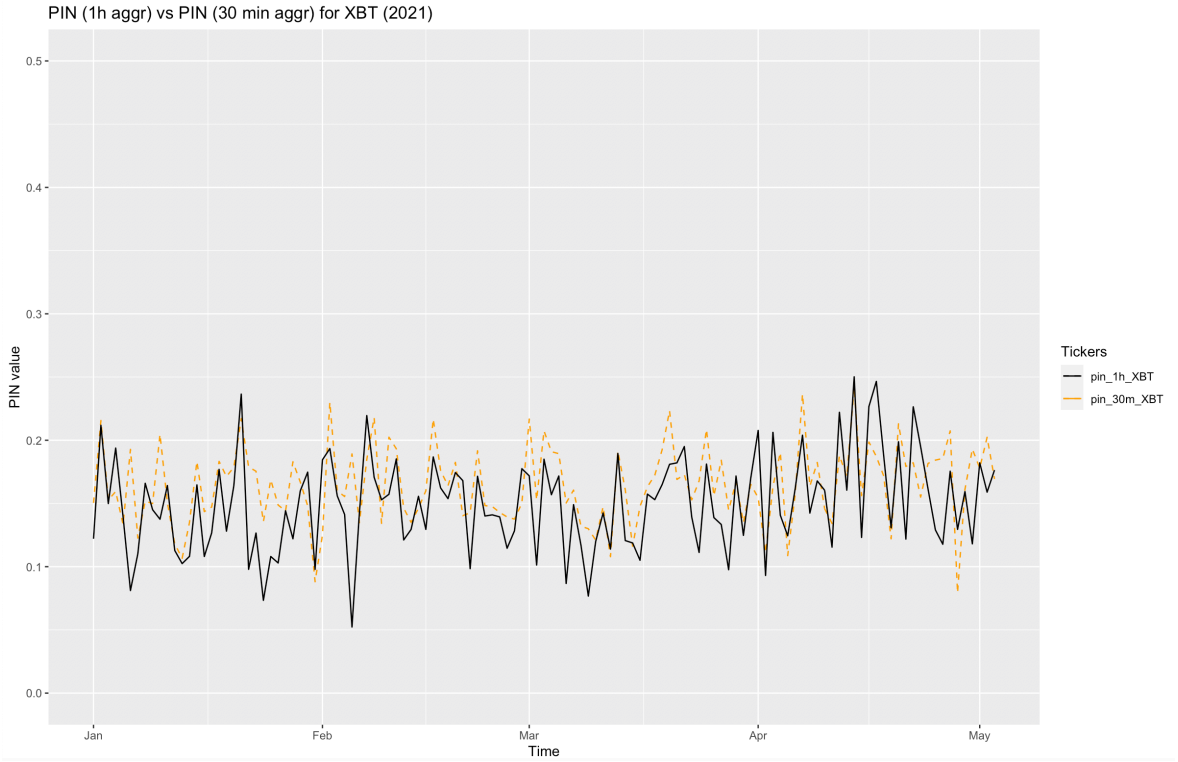


Figure 6: PIN (1h) vs PIN (30 min) for ETH (2021)

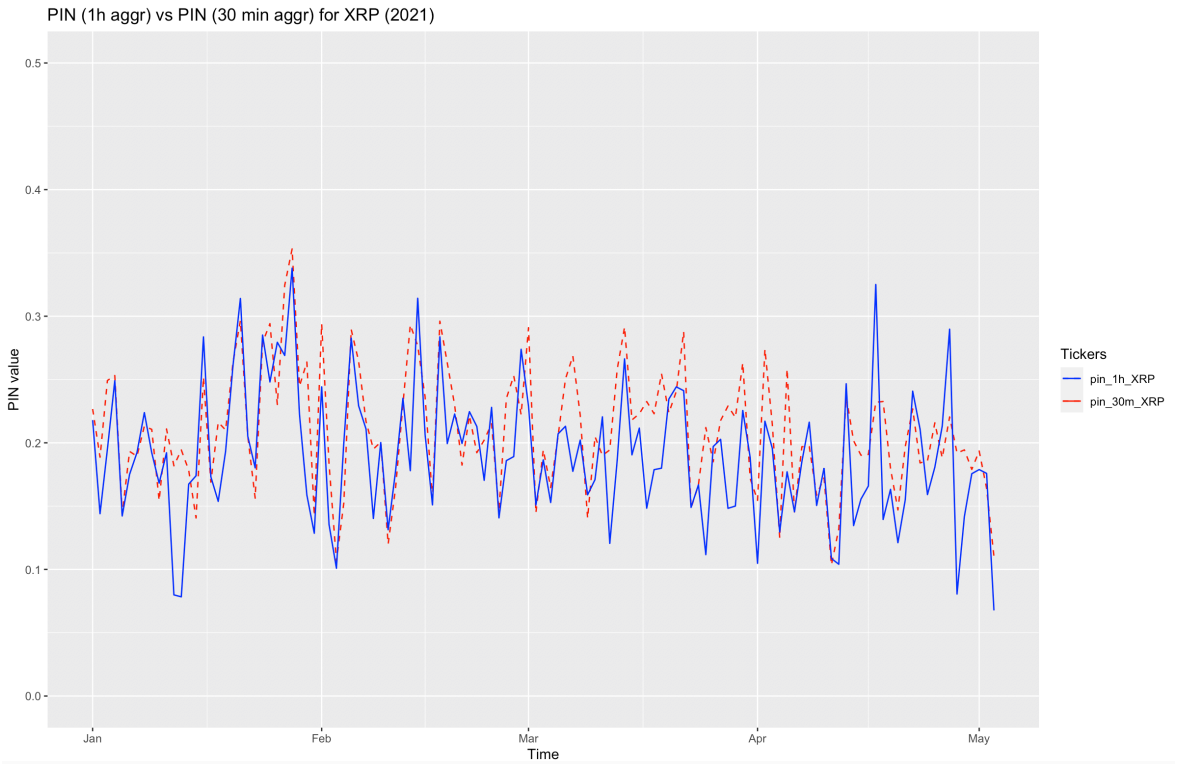
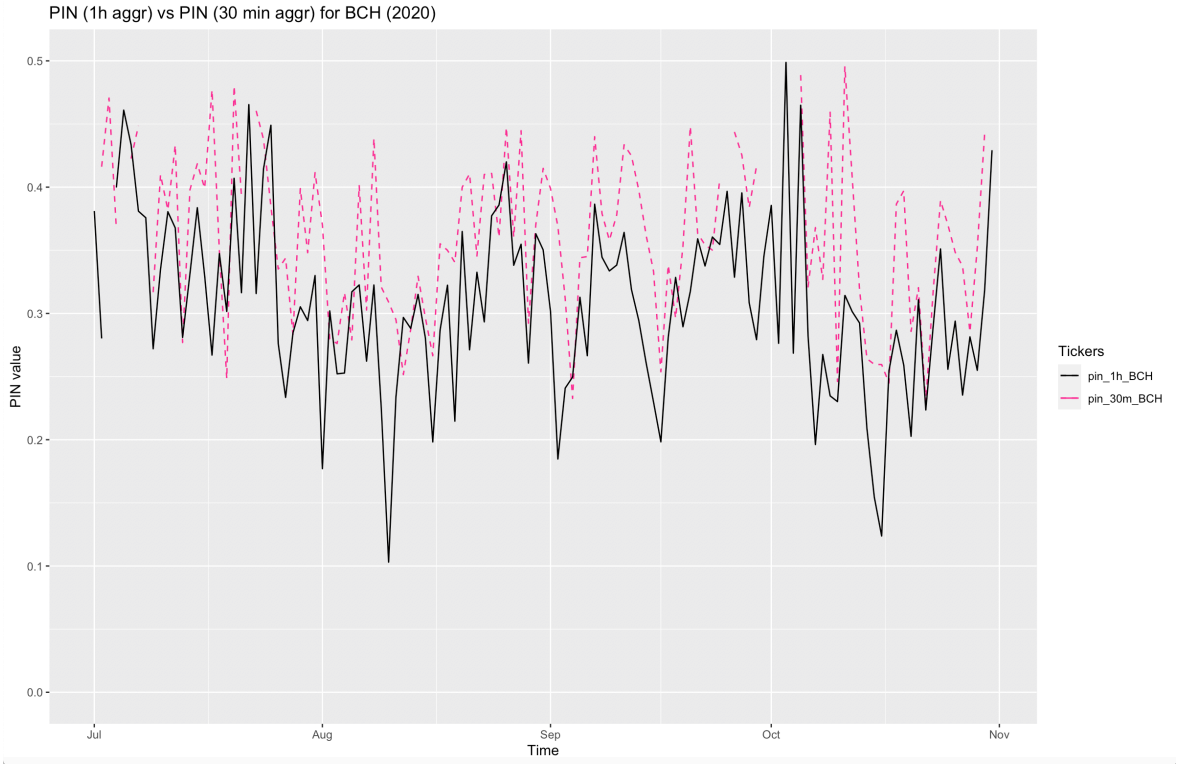


Figure 7: PIN (1h) vs PIN (30 min) for ETH (2021)



The graphs above show that PIN (1h) and PIN (30 min) provide more or less the same results, while the latter is more computationally intensive. However, intuitively, it is worth investigating aggregation at even lower level, such as 5 or 1 minutes, in the future research. This might yield interesting results due to the nature of HFT environment where trades occur at milliseconds.

6.4 PIN Summary

Overall, all types of aggregation have the same features, while lower aggregation cases PIN(2h), PIN (1h) and PIN(30 min) are much smoother and almost identical. Daily aggregation specification PIN(1d) is very infrequent and volatile as it uses much less informative input (for more illustrative statistical summary see Appendix).

Figure 8: PIN (1d) vs PIN (2h) vs PIN (1h) vs PIN (30m) for XBT (2021)

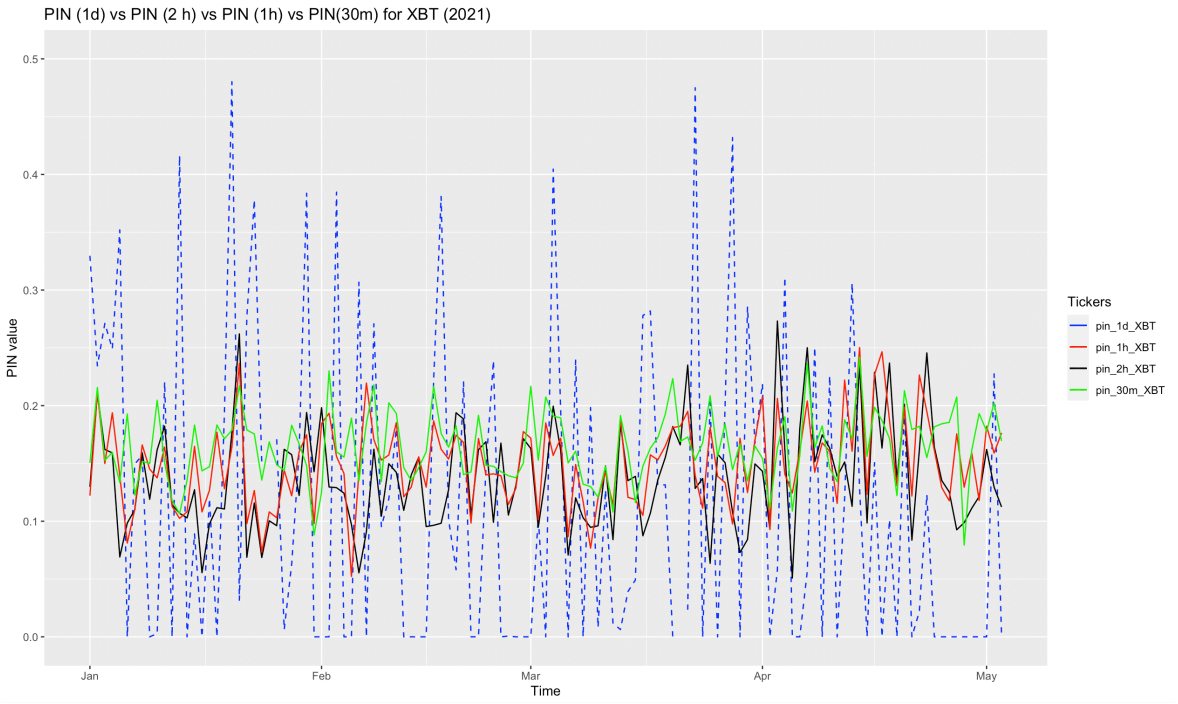
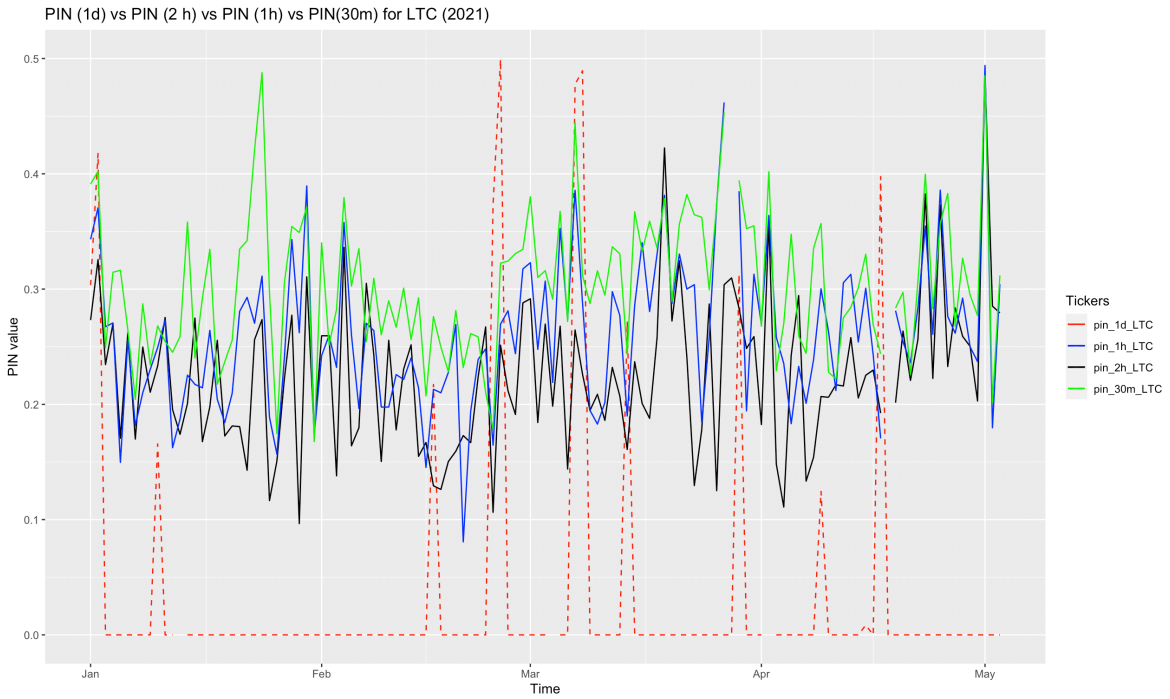


Figure 9: PIN (1d) vs PIN (2h) vs PIN (1h) vs PIN (30m) for LTC (2021)



The application of different PIN model specifications can be summarrized in the following stylized facts:

1) Illiquid cryptocurrencies have higher PIN values, on average

2) Higher aggregation level PIN is not monotonous and has a "saw" shape. It has large proportion of zero values due to the absence of enough information in such a high level of aggregation. However, in case of presence of private information event it provides with higher PIN value, compared to lower levels of aggregation

3) Lower levels of aggregation of Buy and Sell trades appear to contain much more information and make PIN model more efficient in spotting private information. In contrast to higher order of aggregation, they show permanently non-zero PIN series which are smooth and less volatile.

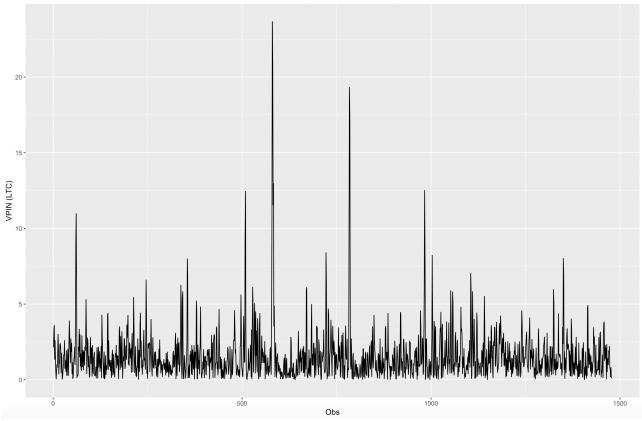
7 VPIN empirics

Another model we apply is VPIN. The main problem with this metric lies in that it is measured in volume universe and it might be confusing to move from volume to time paradigm. Practically, VPIN is measured on the rolling-over basis, which is it estimated via 50 buckets and then is recalculated by dropping the first bucket and including a new one and so on. For instance, for LTC there is the following rolling-window VPIN:

Table 2: Rolling-window for LTC (Dec. 2018)

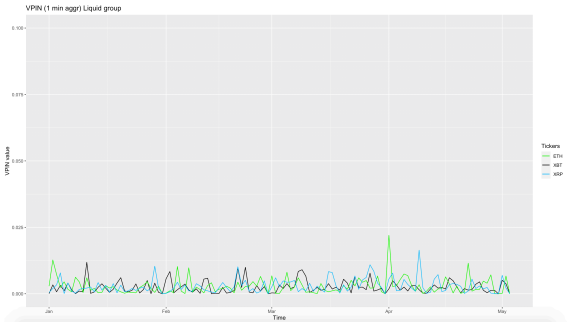
Obs	VPIN	Initial bucket	Final bucket
1	0.02	1	50.00
2	0.03	2	51.00
3	0.04	3	52.00
4	0.01	4	53.00
5	0.03	5	54.00
6	0.02	6	55.00
...	...	...	...

Figure 10: VPIN (%) Rolling-window for LTC (Dec. 2018)

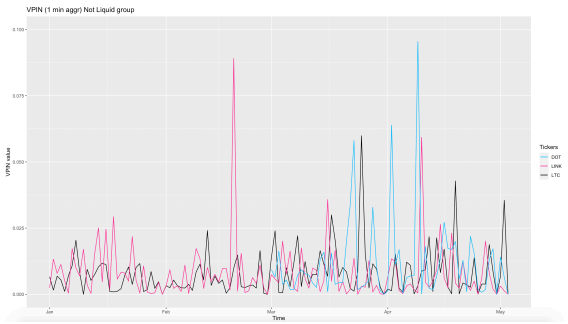


Still, this approach cannot be intuitively moved to time universe and we can hardly obtain relevant daily VPIN values. That is why, in this paper we fill N buckets (via VBS value) and fix the date of formation. The timing might be not homogenous as days differ in liquidity and for some it might take even several days to complete N buckets. However, we ignore that and consider Vpin value for a particular day with respect to the starting point which is the beginning of the day.

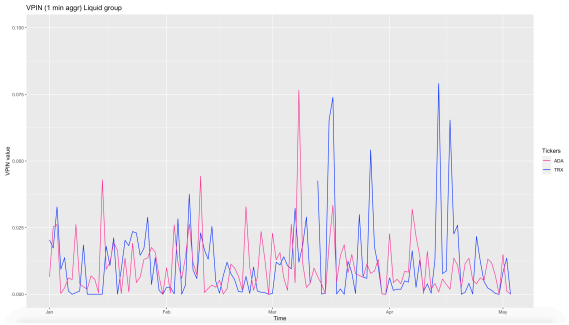
Figure 11: VPIN (1 min aggr, 50 buckets) Liquid group vs Not liquid group (2021)



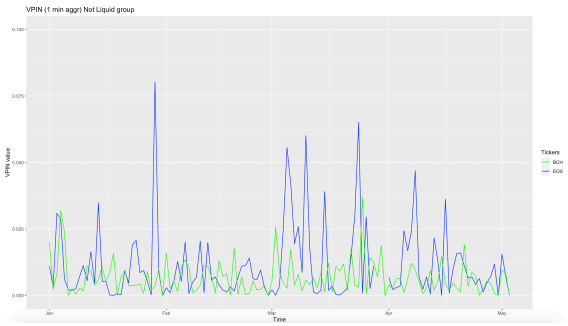
(a) Liquid group (XBT, XRP, ETH)



(b) Not Liquid group (DOT, LINK, LTC)



(c) Liquid group (ADA, TRX)



(d) Not Liquid group (BCH, EOS)

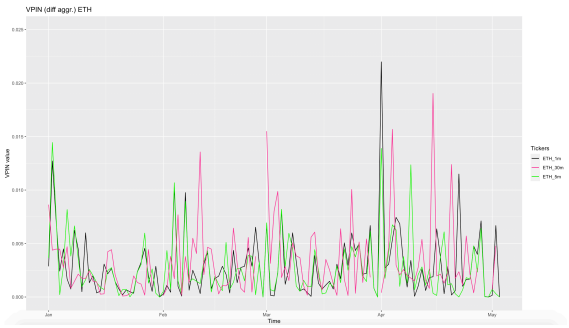
The first aspect we observe is extremely low values of VPIN, compared to PIN. Throughout all four considered years any cryptocurrency's VPIN hardly exceeded 10% probability. Such interesting pattern can be explained as a result of large proportion of noise traders. It can be inferred even from estimated PIN parameters from the previous section that cryptocurrency markets are very deep and such metrics as VPIN that is primarily based on order imbalance might fail to capture the presence of private information in its absolute value. However, its relative change might also be a good indicator of some information-linked activity. Another reason behind such moderate values may be directly related to a level of aggregation that is rather too large or small.

Still, the obtained results are consistent with PIN metric and show that Not Liquid cryptocurrencies appear to have larger VPIN values, on average.

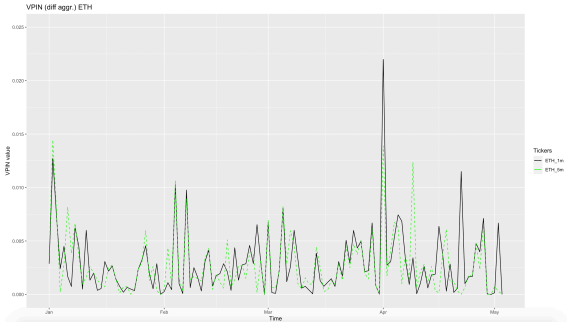
7.1 VPIN series (different level of aggregation comparison)

In order to analyze the relevancy of VPIN low absolute value to the level of aggregation, we compare different VPIN specifications.

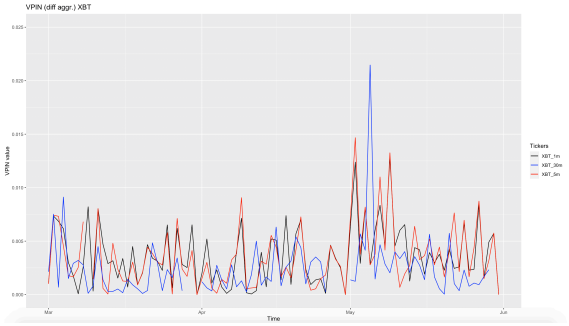
Figure 12: Different levels of aggregation for VPIN on four cryptocurrencies (2018-2021)



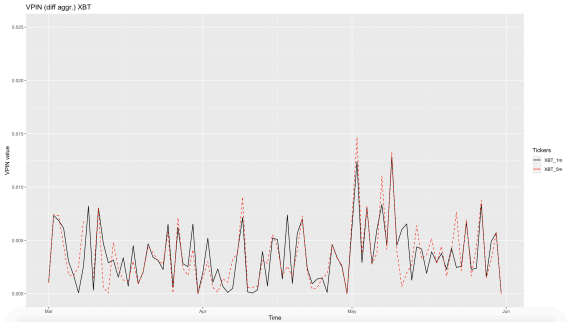
(a) VPIN-s for ETH (2021)



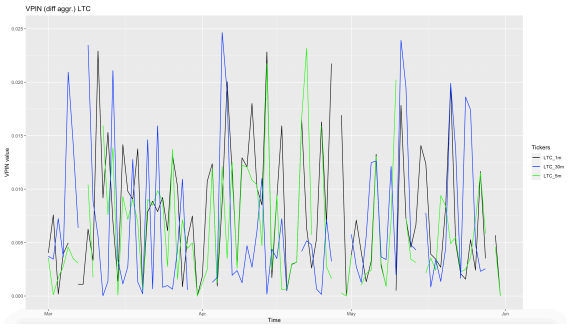
(b) VPIN 1 min vs VPIN 5 min for ETH (2021)



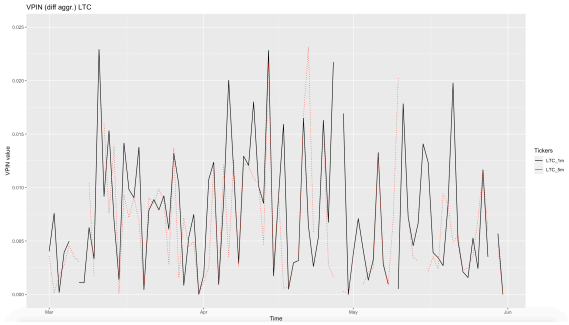
(c) VPIN-s for XBT (2020 Mar-Jun)



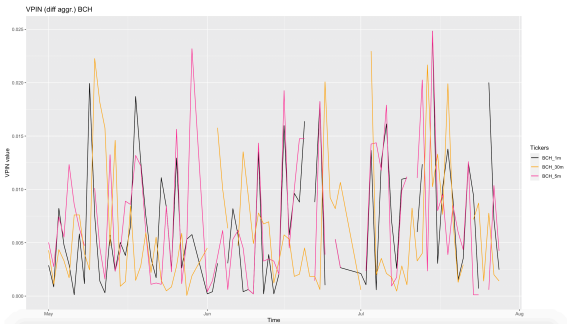
(d) VPIN 1 min vs VPIN 5 min for XBT (2020 Mar-Jun)



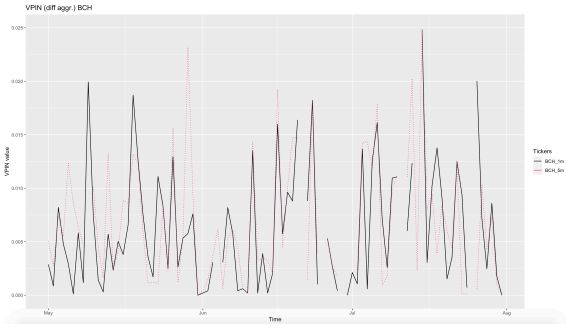
(e) VPIN-s for LTC (2019 May-Aug)



(f) VPIN 1 min vs VPIN 5 min for LTC (2019 May-Aug)



(a) VPIN-s for BCH (2018 May-Aug)



(b) VPIN-s (1 min vs 5 min) for BCH (2018 May-Aug)

From the graphs above it is obvious that 5 min and 1 min VPIN-s are almost identical. Higher level of aggregation, which is 30 min, makes VPIN series less smooth and more volatile. This is exactly the same pattern we observed in case of PIN, where series with higher level of aggregation (1 day) were sharper, subject to extensive fluctuations with "saw" shape in contrast to PIN series with lower trading data aggregation (2 hours, 1 hour or 30 min) which appeared to be smoother and more monotonous. Thus, for VPIN the usage of Buy and Sell trades aggregation at order higher than 1 or 5 minutes is useless and results into not interpretable output, lacking any underlying intuition. As we deal with high frequency data, in future researches VPIN should be tested on lower level of aggregation which are 30 seconds or even 5 seconds as in HFT environment insider traders may hide their trades in these small time periods.

To sum up, we can outline the following **stylized facts** about VPIN metrics:

1) As in case of PIN, illiquid stocks tend to have higher VPIN values than liquid ones, on average

2) As cryptocurrency markets are very liquid, close levels of trading data aggregation result in almostly the same VPIN values, e.g. VPIN 1 min = VPIN 5 min

3) Due to HFT environment in cryptocurrency markets, high levels of aggregation such as 30 minutes and more are not efficient as this kind of data contains a lot of noise and it is hard to spot insider trading via order imbalance. Thus, 1 minute or even intervals of several seconds should be used.

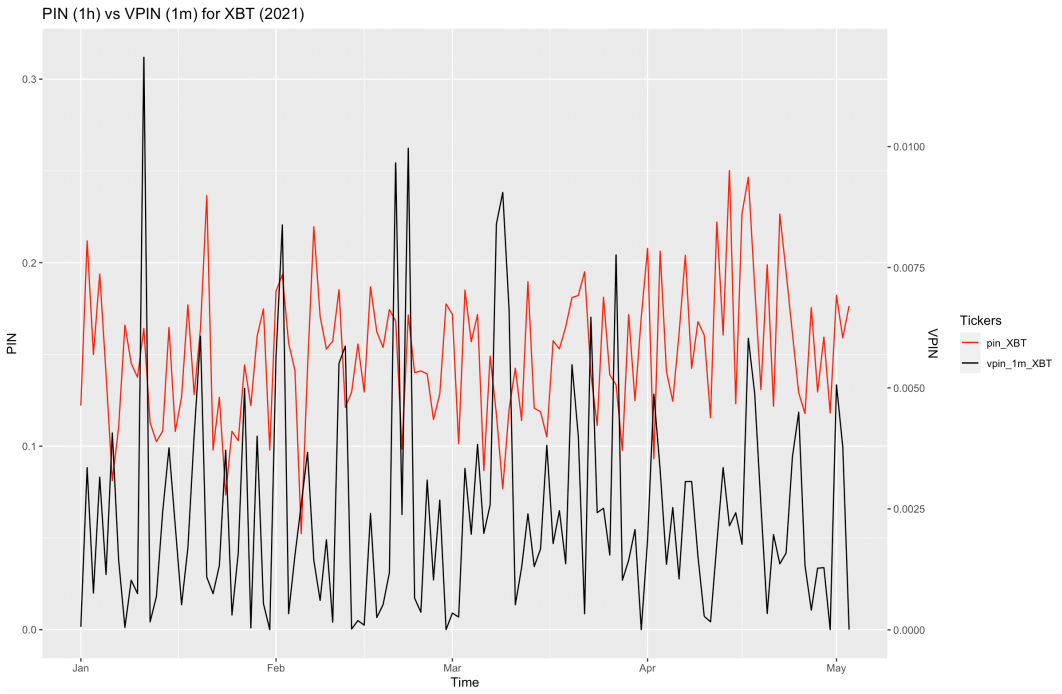
8 PIN vs VPIN

In cryptocurrency HFT environment it appears to be difficult to compare PIN and VPIN directly due to extreme difference in their absolute values.

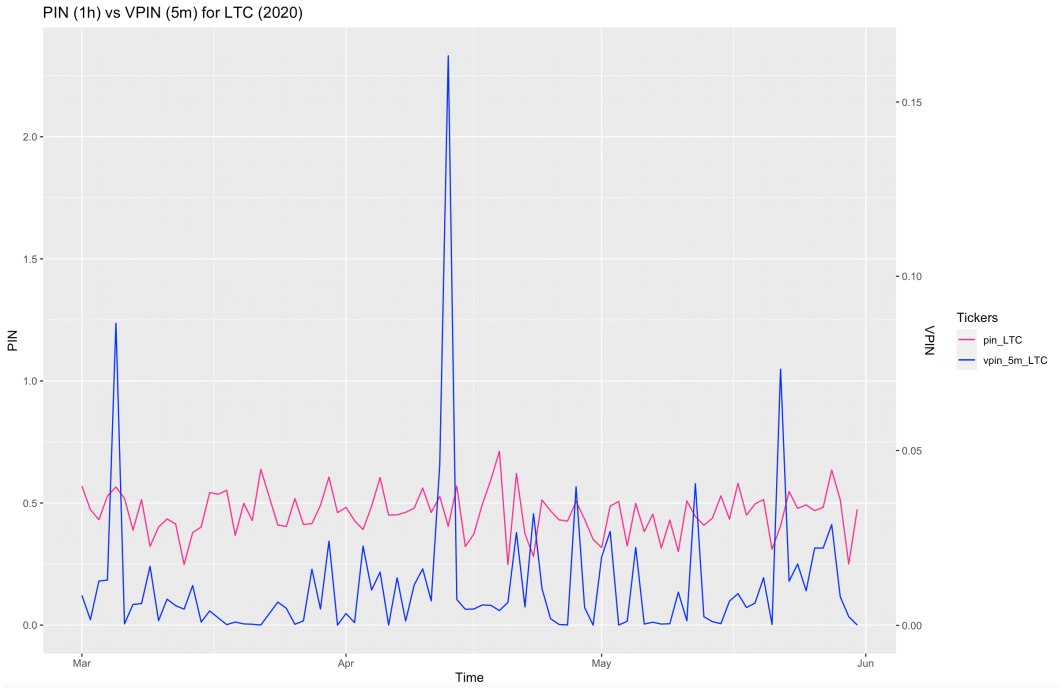
Table 3: PIN vs VPIN summary (2018-2021)

Ticker	Avg. VPIN (1m)	Avg. VPIN (5m)	Avg. PIN (1h)	Avg. PIN (2h)	Avg. PIN (1d)
XBT	0.00395	0.00400	0.18257	0.17064	0.09421
TRX	0.01201	0.01063	0.38563	0.33165	0.02695
XRP	0.00811	0.00763	0.28948	0.25476	0.03398
ADA	0.01112	0.01106	0.37067	0.31753	0.03123
EOS	0.01220	0.01165	0.34517	0.30597	0.02236
LTC	0.01012	0.01028	0.36655	0.31632	0.02815
ETH	0.00535	0.00506	0.21157	0.18841	0.02869
BCH	0.00986	0.00951	0.35336	0.30562	0.02811
LINK	0.00845	0.00788	0.25194	0.09931	0.01121
DOT	0.02089	0.01964	0.12451	0.122813	0.00243

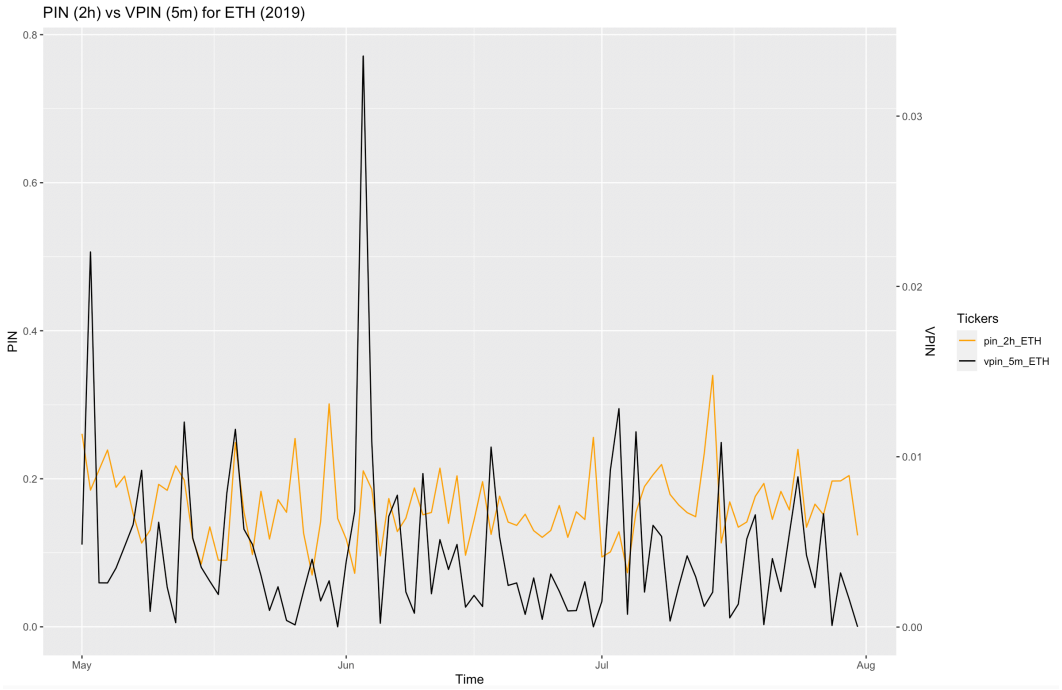
Throughout four years VPIN-s appear to be much lower, on average, than any PIN specification due to the reasons, identified in the previous section. However, we might consider the changes in VPIN as also a good indicator of the change in the presence of insider activity and compare the direction of these alterations to the ones in PIN.



(a) PIN (1h) vs VPIN (1m) for XBT (2021)



(a) PIN (1h) vs VPIN (5m) for LTC (2020 Mar-Jun)



(a) PIN (2h) vs VPIN (5m) for ETH (2019 May-Aug)

The figures above show that although they differ in absolute values, PIN and VPIN metrics are mainly aligned in their behaviour. They have many identical peaks, increase and decrease simultaneously. This supports our supposition that VPIN is still efficient as despite its extremely small values due to extreme liquidity in cryptocurrency, it still can spot insider activity via direction of change in its value.

9 Future research potential

9.1 Positive Correlation

Throughout all four years in all cryptocurrencies there is strong positive correlation, for instance, in 2021 (for other years see Appendix):

Ticker	Mean	Median	Max	Min
XBT	0.903236780439141	0.927285710279863	0.994232508134612	0.326830739374715
TRX	0.471349874710302	0.508133566320679	0.986947182106762	-0.355750913110039
XRP	0.885642578638547	0.943933865321629	0.998558966784805	0.0266919818425183
ADA	0.607637916246464	0.636195161146313	0.975495967744776	-0.159348278829358
EOS	0.630896538110012	0.686472427434494	0.994884957476927	-0.0853756300182049
LTC	0.615041816467201	0.691750339500402	0.995539683734783	-0.200081834457365
ETH	0.884448905512858	0.922683319747954	0.996991691643646	0.407373866084212
BCH	0.464198350184648	0.531088457012869	0.985497276840607	-0.545111663026552
LINK	0.728682922916421	0.782797684377285	0.99049398644141	0.136962389552898
DOT	0.573346956422284	0.638311364768536	0.987038230901411	-0.123871030738398

Thus, this might imply that PIN (EHO) specification might be not efficient for probability of insider trading estimation, since it simply does not allow positive correlation between Buy and Sell trades.

$$cov(B,S) = \alpha^2\mu^2\delta(\delta - 1) \leq 0 \text{ as } 0 \leq \delta \leq 1$$

As an alternative we can use PIN model's modification, called *Adjusted PIN* which was introduced in Duarte and Young (2009), observing similar strong positive correlation between Buy and Sell trades in the stock market.

9.2 Adjusted PIN

This model is a nested version of PIN (EHO) model. It was introduced by Duarte and Young (2007) and has similar but modified framework. As in the original model, there are three types of traders, and with probability α there is a private information event, which can be either negative or positive with underlying probabilities δ and $(1 - \delta)$, respectively. However, informed traders are no longer homogenous, i.e. they

buy and sell at different rates: μ_b and μ_s. Moreover, there is an event of symmetric order flow (with probability θ if private signal is present and probability θ' in case it is absent), such that there are additional Sell (Δ_s) and Buy (Δ_b) orders at the same time. This adjustment reflects empirically observed positive correlation between Buy and Sell orders, which original PIN model fails to account for.

Figure 17: Adjusted PIN model Prob. tree

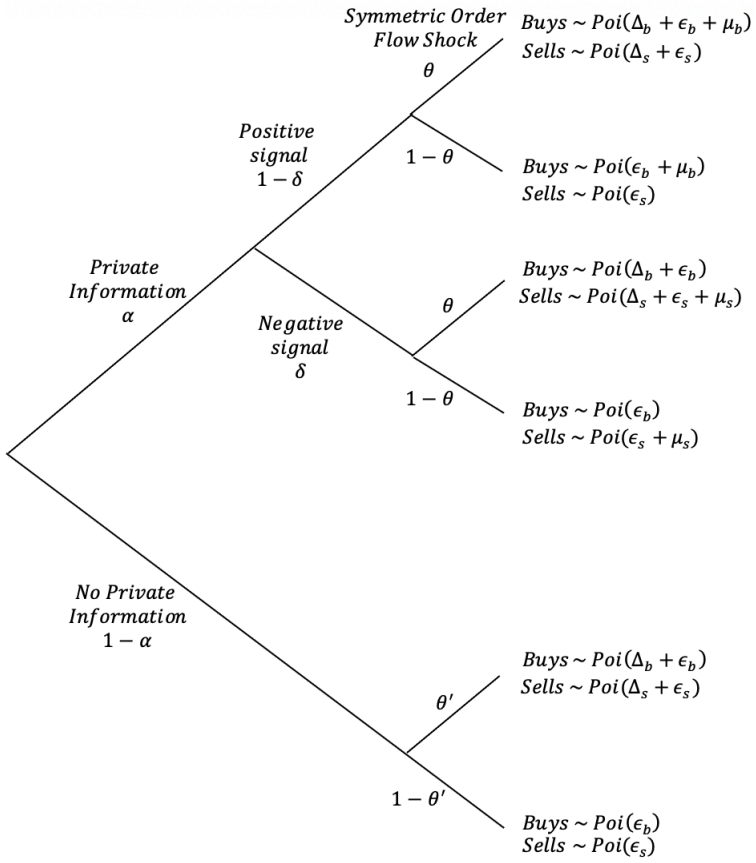


Figure **17: Adjusted trading process tree.** This diagram represents the trading mechanics, where α, δ, θ, θ', μ_b, μ_s, ϵ_b, ϵ_s, Δ_b and Δ_s stay for probabilities of private information event, of negative signal and of symmetric order flow in case of private event and its absence, rate of informed buy and sell operations, rates of noisy buy and sell operations and symmetric buy and sell rates, respectively.

Using homogenous Poisson processes, the following Likelihood function is derived:

$$\begin{aligned}
L(\Theta|B,S) =&(1-\alpha)(1-\theta)e^{-\epsilon_b}\frac{\epsilon_b^B}{B!}e^{-\epsilon_b}\frac{\epsilon_s^S}{S!} \\
&+(1-\alpha)\theta e^{-(\epsilon_b+\Delta_b)}\frac{(\epsilon_b+\Delta_b)^B}{B!}e^{-(\epsilon_s+\Delta_s)}\frac{(\epsilon_s+\Delta_s)}{S!} \\
&+\alpha(1-\theta')(1-\delta)e^{-\epsilon_b}\frac{\epsilon_b^B}{B!}e^{-(\epsilon_s+\Delta_s)}\frac{(\epsilon_s+\Delta_s)^S}{S!} \\
&+\alpha\theta'(1-\delta)e^{-(\epsilon_b+\Delta_b)}\frac{(\epsilon_b+\Delta_b)^B}{B!}e^{-(\mu_s+\epsilon_s+\Delta_s)}\frac{(\mu_s+\epsilon_s+\Delta_s)^S}{S!} \\
&+\alpha(1-\theta')\delta e^{-(\mu_b+\epsilon_b)}\frac{(\mu_b+\epsilon_b)^B}{B!}e^{-\epsilon_s}\frac{\epsilon_s^S}{S!} \\
&+\alpha\theta\delta e^{-(\mu_b+\epsilon_b+\Delta_b)}\frac{(\mu_b+\epsilon_b+\Delta_b)^B}{B!}e^{-(\epsilon_s+\Delta_s)}\frac{(\epsilon_s+\Delta_s)^S}{S!}
\end{aligned}$$

where $\Theta = (\alpha, \delta, \theta, \theta', \mu_b, \mu_s, \epsilon_b, \epsilon_s, \Delta_b, \Delta_s)$ are no news, bad news, probability of symmetric buy and sell trades, given there is private signal, probability of symmetric buy and sell trades, given there is no private signal, insider's buy and sell trading rates, noise traders' buy and sell trading rates, additional buy and sell trading rates in case of symmetric trading event, respectively, while B and S are total Buy and Sell operations per day.

As before we formulate the maximization problem for t periods as:

$$V = \prod L(\Theta|B,S) = \sum logL(\Theta|B,S)$$

Formula for Adjusted PIN is the same ratio of expected insider trading order flow to total order flow (nested PIN formula):

$$Adj\ PIN = \frac{\alpha \times (\delta \times \mu_s + (1-\delta) \times \mu_b)}{\alpha \times ((1-\delta) \times \mu_b + \delta \times \mu_s) + (\Delta_b + \Delta_s) \times (\alpha \times \theta' + (1-\alpha) \times \theta) + \epsilon_s + \epsilon_b}$$

Introduction of symmetric buy and sell trading orders event (θ and θ') in Adjusted PIN model helps to reflect the empirical positive correlation between buy and sell orders, which is always negative in the original PIN set-up.

9.3 Introducing modified version of Yang and Zhang (2012) algorithm, fitted for Adjusted PIN model

Being nested model of PIN, Adjusted PIN inherits all the computational difficulties, discussed in previous sections. One of them is the choice of initial parameters in ML maximization procedure, which was proven to be crucial by Yang and Zhang (2012) in case of PIN. Thus, we create our own initial parameters choice approach, using intuition, similar to the one of Yang and Zhang (2012).

We use the method of moments approach to evaluate μ_b and μ_s:

$$\mathbb{E}(B) = \alpha\theta\Delta_b + (1-\alpha)\theta'\Delta_b + \alpha(1-\delta)\mu_b + \epsilon_b$$
$$\mathbb{E}(S) = \alpha\theta\Delta_s + (1-\alpha)\theta'\Delta_s + \alpha\delta\mu_s + \epsilon_s$$

Using sample values, we can infer from equations above that:

$$\mu_b = \frac{\bar{B} - \epsilon_b - \Delta_b(\alpha\theta + (1-\alpha)\theta')}{\alpha(1-\delta)}$$
$$\mu_s = \frac{\bar{S} - \epsilon_s - \Delta_s(\alpha\theta + (1-\alpha)\theta')}{\alpha\delta}$$
$$\epsilon_b = \gamma\bar{B} \qquad \epsilon_s = \gamma\bar{S}$$
$$\Delta_b = ? \qquad \Delta_s = ?$$

How to choose initial values of additional buy and sell orders (Δ_b and Δ_s)?

The main problem arrises with the choice of Δ_b and Δ_s. We cannot set them theoretically, still there are several ways we can define them (intuitively):

1) Difference between the realized number of Buy (Sell) orders and the mean Buy (Sell) orders within a trading day:

$$\Delta_{b,i} = \bar{B} - B_i \text{ and } \Delta_{s,i} = \bar{S} - S_i$$

2) As in case of ϵ_b and ϵ_s set them as fractions of mean number of Sell and Buy orders (computationally intensive):

$$\Delta_{b,i} = \eta\bar{B} \text{ and } \Delta_{s,i} = \eta\bar{S}$$

where $\eta \in (0.1,0.3,0.5,0.7,0.9)$ and $\mu_b, \mu_s \geq 0$

3) The simplest approach is let them be defined in the range $[0; +\infty)$, such that $\mu_b \geq 0$ and $\mu_s \geq 0$

$\alpha,\delta,\theta,\theta',\gamma$ (and η if 2-nd approach is used) take one of equally-distanced values (0.1,0.3,0.5,0.7,0.9) at a time (there are 3125 (15625) possible combinations), $\bar{B}$ and $\bar{S}$ are estimators of expected values of B and S respectively. Thus, in the same fashion as in the Yang and Zhang (2012), we arrive at the following procedure: firstly, we run maximization for the 3125 (15625) sets of the possible values, excluding those where μ_s or μ_b is negative ($\mu_s < 0$ or $\mu_b < 0$). Secondly, if all solutions are on the boundary, we choose the one with the highest value of the Likelihood function, otherwise we exclude them and choose the one among non-boundary, using the same approach.

9.4 Factorization for Adjusted PIN

Another inherited problem is traditional factorisation that does not allow to estimate high number of trading orders which we face in liquid markets. As there is no modified likelihood function present (such as Lin and Ke (2011) for PIN) this paper suggests the following substitution[3]:

$$e^{-\epsilon_b}\frac{\epsilon_b^B}{B!} \approx e^{-\epsilon_b + B\cdot ln(\epsilon_b) - \sum_{i=1}^{B} ln(i)}$$

We apply this trick to all parts of the Likelihood function and as result obtain the

[3]For complete derivation see Appendix

following modified Likelihood for Adjusted PIN:

$$
\begin{aligned}
L(\Theta|B,S) =& (1-\alpha)(1-\theta)e^{-\epsilon_b + B\cdot ln(\epsilon_b) - \sum_{i=1}^{B} ln(i)} \cdot e^{-\epsilon_s + S\cdot ln(\epsilon_s) - \sum_{i=1}^{S} ln(i)} \\
&+(1-\alpha)\theta e^{-\epsilon_b - \Delta_b + B\cdot ln(\epsilon_b + \Delta_b) - \sum_{i=1}^{B} ln(i)} \cdot e^{-\epsilon_s - \Delta_s + S\cdot ln(\epsilon_s + \Delta_s) - \sum_{i=1}^{S} ln(i)} \\
&+\alpha\theta'(1-\delta)e^{-\epsilon_b + B(\epsilon_b) - \sum_{i=1}^{B} ln(i)} \cdot e^{-\mu_s - \epsilon_s + S\cdot ln(\mu_s + \epsilon_s) - \sum_{i=1}^{S} ln(i)} \\
&+\alpha(1-\theta')\delta e^{-\mu_b - \epsilon_b + B\cdot ln(\mu_b + \epsilon_b) - \sum_{i=1}^{B} ln(i)} \cdot e^{-\epsilon_s + S\cdot ln(\epsilon_s) - \sum_{i=1}^{S} ln(i)} \\
&+\alpha\theta\delta e^{-\mu_b - \epsilon_b - \Delta_b + B\cdot ln(\mu_b + \epsilon_b + \Delta_b)} \cdot e^{(-\epsilon_s - \Delta_s + S\cdot ln(\epsilon_s + \Delta_s) - \sum_{i=1}^{S} ln(i))}
\end{aligned}
$$

10 Conclusion

This paper presents the stylized facts on probability of informed trading in cryptocurrency markets. One of the main findings is that less liquid cryptocurrencies tend to have higher probability of insider trading than liquid ones (both PIN and VPIN are higher, on average), which is consistent with Easley et al (1996) findings for stock markets. Less frequently traded cryptocurrencies have lower proportion of noise traders, compared to liquid ones, thus, appear to be more risky as there is higher probability that a trade is performed by an insider. Although liquid cryptocurrencies tend to have higher probability of informed trading due to larger number of public events, related to them, this effect is offset by greater proportion of noise traders, reflecting higher degree of market depth.

Both VPIN and PIN metrics in case of high level of aggregation have a "saw" shape, fluctuate a lot and have large proportion of zero-values. This is a consequence of small amount of information due to high liquidity in cryptocurrency markets which makes such a high order of aggregation too noisy to spot insider trading effectively. Thus, it is much more efficient to use lower level of aggregation which will contain much more accurate information on trading order imbalance. This is supported by both empirical PIN and VPIN estimates that at lower order of aggregation appear to be smooth, less volatile and persistently non-zero.

VPIN metric was found to be extremely low for all types of cryptocurrencies. The main potential reason behind this phenomenon are deep cryptocurrency markets that contain a lot of noise traders, so it is more difficult to spot informed trading activity. Another potential reason is HFT environment that might imply that standard 1 min aggregation is too high and even lower aggregation order is required.

Still, VPIN metric was found to be aligned with PIN in terms of direction of change, meaning that although it is very low in absolute value the direction of its change is a good indicator of some potential appearance of information-linked trades in the market.

Moreover, some levels of aggregation resulted into almost identical PIN and VPIN

values [for instance, (VPIN 1 min and VPIN 5min) or (PIN 1h and PIN 2h)]. As lower order of aggregation is usually more computationally intensive, a close, but higher level of aggregation can be used to estimate PIN or VPIN metrics for some potential trading activities where estimation speed is vital. However, this aspect requires further investigation.

Another empirical fact which mimics the evidence for stock markets is strong positive correlation between Buy and Sell trades, which implies that PIN model might not be efficient as it theoretically does not allow for positive correlation. Thus, this study proposes to use Adjusted PIN model by Duarte and Young (2007) in future research and introduces some theoretical innovations to its estimation.

Nevertheless, there is high potential for further investigation. Research papers, criticising some empirical pattern of PIN and VPIN metric in the stocks markets can be replicated. For instance, as for VPIN, Anderson and Bondarenko (2014) claimed that VPIN value depends heavily on the starting point which supports our concerns about difficulties with translation from volume to time universe, discussed in this paper. Furthermore, as we deal with HFT environment lower order of aggregation such 5 min and 1 min aggregation for PIN and even several seconds unification for VPIN should be studied more carefully. Finally, the choice of optimal number of buckets should be analyzed as this research used the conventional 50, proposed by Easley et al (2012) for the stocks markets.

11 Bibiliography

[1] Aktas, Nihat, de Bodt E., Declerck F., and Van Oppens H. (2007), The PIN anomaly around M & A announcements, Journal of Financial Markets 10, 169—191.

[2] Amihud, Yakov (2002), Illiquidity and stock returns: cross section and time-series effects, Journal of Financial Markets 5, 31-56

[3] Andersen, T.G., and O. Bondarenko (2014), VPIN and the Flash Crash, Journal of Financial Markets, 17, pp. 1-46.

[4] Collin-Dufresne P., and Fos V. (2012), Do prices reveal the presence of informed trading?, Journal of Finance Forthcoming.

[5] Duarte, Jefferson, and Young L. (2009), Why is PIN priced?, Journal of Financial Economics 91, 119—138.

[6] Easley D., Hvidkjaer S., and O'Hara M. (2010), Factoring information into returns. Journal of Financial and Quantitative Analysis.

[7] Easley D., Kiefer N., O'Hara M., and Paperman J. (1996), Liquidity, information, and infrequently traded stocks, Journal of Finance 51, 1405- 1436.

[8] Easley, D., López de Prado, M., O'Hara, M., (2012). Flow toxicity and liquidity in a high-frequency world. Review of Financial Studies 25, 1457–1493.

[9] Easley D., O'Hara M. (1987), Price, trade size, and information in securities markets, Journal of Financial Economics 19, 69-90.

[10] Gan Q., Wei W.C., Johnstone D. (2015) A faster estimation method for the probability of informed trading using hierarchical agglomerative clustering. Quantitative Finance, 15(11):1805–1821.

[11] Glosten L. and P. Milgrom (1985), Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, Journal of Financial Economics 13, 71-100.

[12] Hellwig M. (1980), "On the Aggregation of Information in Competitive Markets," Journal of Economic Theory, 22, 477-498.

[13] G. Kuzmin (2020)

[14] Kyle A. S. (1985), Continuous auctions and insider trading, Econometrica 53, 1315—1335.

[15] H.-W. W. Lin and W.-C. Ke (2011), A computing bias in estimating the probability of informed trading. Journal of Financial Markets, 14(4), 625–640.

[16] E. R. Odders-White and M.J. Ready (2008), The probability and magnitude of information events, Journal of Financial Economics, 87, 227–248.

12 Appendix

12.1 Derivation of factorisation for Adjusted PIN

Let us consider expression:

$$e^{-\epsilon_b}\frac{\epsilon_b^B}{B!}$$

Taking log:

$$-\epsilon_b + Bln\epsilon_b - lnB!$$

$$\epsilon_b + Bln\epsilon_b - \sum_{t=1}^{B} ln(i)$$

Taking exponent

$$e^{(-\epsilon_b + Bln\epsilon_b - \sum_{t=1}^{B} ln(i))}$$

In the same fashion we perform similar transformation for other parts of Adjusted PIN likelihood function

12.2 Initial Parameters for Adjusted PIN derivation

Using method of moments we can derive Expected Buy and Sell trades:

$$\begin{aligned}\mathbb{E}(B) &= \alpha(1-\delta)[\theta(\Delta_b + \epsilon_b + \mu_b) + (1-\theta)(\epsilon_b + \mu_b)] + \alpha\delta\theta\Delta_b + \\ &+ \alpha\delta\epsilon_b + (1-\delta)\theta'\Delta_b + (1-\alpha)\epsilon_b = \\ &= \alpha\theta\Delta_b + (1-\alpha)\theta'\Delta_b + \alpha(1-\delta)\mu_b + \epsilon_b\end{aligned}$$

$$\mu_b = \frac{\mathbb{E}(B) - \epsilon_b - \Delta_b(\alpha\theta + (1-\alpha)\theta')}{\alpha(1-\delta)}$$

$$\begin{aligned}\mathbb{E}(S) &= \alpha(1-\delta)[\theta(\Delta_s+\epsilon_s)+(1-\theta)\epsilon_s)]+\alpha\delta[\theta(\Delta_s+\epsilon_s+\mu_s)+\\ &+(1-\theta)(\epsilon_s+\mu_s)]+(1-\alpha)\theta'\Delta_s+(1-\alpha)\epsilon_s\\ &=\alpha\theta\Delta_s+(1-\alpha)\theta'\Delta_s+\alpha\delta\mu_s+\epsilon_s\end{aligned}$$

$$\mu_b = \frac{\mathbb{E}(B)-\epsilon_b-\Delta_b(\alpha\theta+(1-\alpha)\theta')}{\alpha(1-\delta)}$$

12.3 Tables and graphs

Table 4: PIN (1d) Summary 2018-2021

Ticker	Average PIN (1d)	Median PIN (1d)	St. dev PIN (1d)
XBT	0.09421	0.00000	0.14391
TRX	0.02695	0.00000	0.09702
XRP	0.03398	0.00000	0.11287
ADA	0.03123	0.00000	0.10611
EOS	0.02236	0.00000	0.09352
LTC	0.02815	0.00000	0.10455
ETH	0.02869	0.00000	0.10066
BCH	0.02811	0.00000	0.10755
LINK	0.01121	0.00000	0.06929
DOT	0.00243	0.00000	0.03000

Table 5: PIN (2h) Summary (2018-2021)

Ticker	Average PIN (2h)	Median PIN (2h)	St. dev PIN (2h)
XBT	0.17064	0.16668	0.05887
TRX	0.33165	0.34152	0.18127
XRP	0.25476	0.24798	0.08505
ADA	0.31753	0.30902	0.10223
EOS	0.30597	0.31944	0.15899
LTC	0.31632	0.30431	0.09890
ETH	0.18841	0.18090	0.06846
BCH	0.30562	0.29578	0.09996
LINK	0.09931	0.00000	0.12865
DOT	0.12281	0.00000	0.23144

Table 6: PIN (1h) Summary (2018-2021)

Ticker	Average PIN (1h)	Median PIN (1h)	St. dev PIN (1h)
XBT	0.18257	0.17960	0.05474
TRX	0.38563	0.40140	0.20534
XRP	0.28948	0.28133	0.09363
ADA	0.37067	0.36166	0.11595
EOS	0.34517	0.36177	0.17329
LTC	0.36655	0.35651	0.10653
ETH	0.21157	0.19809	0.07343
BCH	0.35336	0.34455	0.10760
LINK	0.25194	0.00000	0.28528
DOT	0.12451	0.00000	0.24461

Table 7: Correlation between Buy and Sell trades (2020)

	Ticker	Mean	Median	Max	Min
1	XBT	0.888109962515809	0.926601954504668	0.997530995942261	0.222027890136913
2	TRX	0.357926620109456	0.373371039873025	0.991299993971155	-0.470382531540881
3	XRP	0.678361948171517	0.772677856354026	0.998728377350391	-0.399131100804635
4	ADA	0.455258052872361	0.492447769839042	0.993393583745946	-0.474570116215769
5	EOS	0.212002548811137	0.153430032688595	0.999225132500328	-0.57145525121242
6	LTC	0.501977207603249	0.584569447069158	0.998610906451171	-0.348051562593804
7	ETH	0.833626471778273	0.887461709256994	0.997603547815067	0.1931477537609
8	BCH	0.563505345355012	0.614926626391913	0.992157121903707	-0.313634190172896
9	LINK	0.679630454919847	0.724220613996848	0.993111761218957	0.0523278749912633
10	DOT	0.223042391365012	0.103004961184687	0.976800057584858	-0.530485820300082

Table 8: Correlation between Buy and Sell trades (2019)

	Ticker	Mean	Median	Max	Min
1	XBT	0.890437288516249	0.919129681818169	0.995718980757517	0.42533464129865
2	TRX	0.503561153692183	0.54602958872863	0.993565477189503	-0.451361683134564
3	XRP	0.558301594992166	0.635561444600616	0.99661313824097	-0.501050208318302
4	ADA	0.473494604808316	0.500514782267785	0.987008807636387	-0.39279987977573
5	EOS	0.523153280395536	0.562006114237904	0.989229803923616	-0.237917972114171
6	LTC	0.539113204787021	0.619020942629547	0.989220066972803	-0.407614535763476
7	ETH	0.828414920428858	0.873278778912286	0.994232977801071	-0.00522230106055234
8	BCH	0.532551546519329	0.606873608832653	0.985418146868393	-0.416214448765167

Table 9: Correlation between Buy and Sell trades (2018)

	Ticker	Mean	Median	Max	Min
1	XBT	0.864172287255054	0.886865044003179	0.994142333469577	0.31179477465504
2	TRX	0.492637200630991	0.557817030897828	0.968504924355788	-0.360617497497174
3	XRP	0.612297145107533	0.67065192750489	0.996043469332478	-0.313572118891721
4	ADA	0.558935330239017	0.591920619216227	0.992679385477307	-0.412174677048205
5	EOS	0.638622225174753	0.709167899247407	0.981360106420383	-0.360147214133299
6	LTC	0.508131400292908	0.565452375567043	0.98674087492395	-0.263502170234861
7	ETH	0.684210667214379	0.756313926835361	0.993565960365154	-0.202189772423874
8	BCH	0.578823604740555	0.632504048050193	0.991019978133685	-0.600956426776881